Valutazione Analitica e Strategie Predittive per il "Cyber Crimes Dataset": Un'Analisi Approfondita dei Pattern Geospaziali Sintetici e delle Metodologie di Threat Intelligence

Sezione 1: Analisi Prioritaria - Validazione della Prospettiva Geospaziale (Nazione-Nazione)

Questa sezione risponde direttamente alla prospettiva di analisi indicata come di primario interesse: lo studio degli attacchi tra nazioni basato sugli indirizzi IP. L'analisi fornisce una metodologia end-to-end, dall'arricchimento dei dati grezzi alla visualizzazione avanzata dei flussi di attacco.

1.1 Obiettivo dell'Analisi: La Mappatura dei Conflitti Digitali Simulati

L'obiettivo è trasformare i dati grezzi sugli indirizzi IP contenuti nelle colonne Attacker IP Address e Target IP Address [1] in un'analisi geospaziale strutturata. Questa analisi mira a identificare i pattern di attacco simulati tra diverse nazioni. L'indagine si concentrerà nel rispondere a tre domande fondamentali:

1. Quali nazioni sono le principali *origini* simulate degli attacchi nel dataset?
2. Quali nazioni sono i principali *obiettivi* simulati?
3. Quali sono le *rotte* di attacco (flussi nazione-nazione) più frequenti, e quali sono le loro caratteristiche (es. gravità, tasso di successo)?

1.2 Metodologia di Arricchimento dei Dati: La Trasformazione degli IP in Nazioni

Per mappare gli attacchi a livello nazionale, il primo passo, e forse il più critico, è

l'arricchimento dei dati, un processo che assegna a ciascun indirizzo IP una geolocalizzazione approssimativa.[2]

Fase 1: Scelta della Libreria di Geolocalizzazione

La scelta dello strumento di geolocalizzazione è fondamentale, poiché deve bilanciare accuratezza, velocità e costo, specialmente quando si elaborano decine di migliaia di record.

- Opzione A (Raccomandata): Database Locale (MaxMind GeoLite2)
  L'utilizzo di un database locale è l'approccio più efficiente per l'elaborazione di massa. La libreria geoip2 di Python, utilizzata con i database GeoLite2 gratuiti di MaxMind, è considerata uno standard industriale.2 Richiede il download del file di database (es. GeoLite2-Country.mmdb) 3, ma consente interrogazioni offline estremamente rapide. Altri database simili sono disponibili su Kaggle, come IP2Location LITE 4 o il Global IP Dataset.5
- Opzione B (Alternativa): Servizi Basati su API
  Librerie come ip2geotools 6 o servizi API commerciali come ipstack 9 e WhoisXML API 10 astraggono il database. Sebbene siano facili da implementare, presentano due svantaggi per l'analisi di massa: (1) Richiedono una connessione Internet per ogni lookup, rendendoli ordini di grandezza più lenti; (2) I servizi gratuiti hanno limiti di richiesta stringenti, e quelli a pagamento richiedono chiavi API e un budget.

Fase 2: Implementazione Tecnica in Python/Pandas

Il processo di arricchimento utilizzando l'Opzione A (MaxMind) in un ambiente pandas è il seguente:

1. **Caricamento:** Caricare il dataset "Cyber Crimes Dataset" in un DataFrame pandas.
2. **Inizializzazione:** Inizializzare il database reader di geoip2: reader = geoip2.database.Reader('percorso/GeoLite2-Country.mmdb').[2]
3. **Definizione Funzione:** Creare una funzione Python per gestire il lookup e, cosa cruciale, gli errori. Gli indirizzi IP privati, malformati o non mappabili (es. 127.0.0.1, 192.168.x.x) solleveranno eccezioni come AddressNotFoundError o ValueError.[11] La funzione deve catturare queste eccezioni e restituire un valore nullo (es. None o pd.NA).
   Python
   ```
   import geoip2.database
   import pandas as pd

   # Inizializza il reader (da fare una sola volta)
   try:
      reader = geoip2.database.Reader('GeoLite2-Country.mmdb')
   except FileNotFoundError:
   ```

```
    print("Errore: Database GeoLite2-Country.mmdb non trovato.")
    # Gestire l'errore

def get_country(ip):
    if not ip or pd.isna(ip):
        return pd.NA
    try:
        response = reader.country(ip)
        return response.country.iso_code # Restituisce il codice ISO-2 (es. 'US', 'CN')
    except (geoip2.errors.AddressNotFoundError, ValueError):
        return pd.NA # IP non trovato, privato o malformato
```

4. **Applicazione Efficiente (Ottimizzazione):**
    - *Approccio Naive:* Applicare la funzione a ogni riga: df = df['Attacker IP Address'].apply(get_country). Questo è *lento* poiché ripete i lookup per IP identici.
    - *Approccio Ottimizzato:* Come dimostrato in [11], un metodo molto più performante è mappare solo gli IP unici. Si estraggono tutti gli IP unici da entrambe le colonne (Attacker IP Address e Target IP Address), si applica get_country solo a questa lista unica, si memorizzano i risultati in un dizionario ({ip: country}) e infine si usa pandas.map() per trasmettere i risultati al DataFrame.
    - *Approccio Vettorizzato:* Si può anche utilizzare una libreria specializzata come pandas_maxminddb [13], che è progettata per questa esatta operazione e offre un'API pandas-nativa e potenzialmente parallelizzata.[13]
5. **Risultato:** Il DataFrame conterrà due nuove colonne: Source_Country e Target_Country (contenenti codici ISO-2).

1.3 Analisi Esplorativa (EDA) Geospaziale

Con i dati arricchiti, è possibile eseguire un'analisi esplorativa per quantificare i pattern di attacco:

- **Principali Attaccanti/Vittime:** source_counts = df.value_counts() e target_counts = df.value_counts(). Questo identifica le nazioni più attive come origine e quelle più bersagliate.
- Coppie di Flusso (Rotte): Creare un DataFrame aggregato che quantifichi le rotte di attacco più comuni:
  flow_counts = df.groupby().size().sort_values(ascending=False)
- **Analisi Qualitativa dei Flussi:** Arricchire l'analisi dei flussi aggregando non solo il conteggio, ma anche la gravità media (Attack Severity) e il tasso di successo (Outcome == 'Succeeded') per ciascuna rotta.

1.4 Visualizzazione dei Flussi di Attacco: Dalle Mappe Coropletiche ai

Grafici di Flusso

La visualizzazione è essenziale per comunicare i risultati geospaziali. La libreria plotly [14] offre capacità interattive superiori per questo compito.[15]

- Visualizzazione 1: Mappe Coropletiche (Choropleth Maps)
  Una mappa coropletica è ideale per mostrare il volume totale di attacchi originati o subiti da ciascuna nazione, colorando le nazioni in base a una scala numerica.16
  - **Strumento:** plotly.express.choropleth.[16]
  - **Metodologia:** Si utilizzano i dati aggregati (source_counts o target_counts). Poiché plotly richiede codici paese ISO-3 (es. 'USA') e MaxMind restituisce ISO-2 (es. 'US'), è necessaria una conversione. La libreria pycountry può essere utilizzata per questo passaggio: pycountry.countries.get(alpha_2=iso_code).alpha_3.[3]
  - *Output:* Due mappe distinte: una "Heatmap Globale degli Attaccanti" e una "Heatmap Globale delle Vittime".
- Visualizzazione 2: Mappe di Flusso (Linee su Mappe Geografiche)
  Questa visualizzazione risponde più direttamente alla domanda "attacchi tra nazioni".18 Mostra le connessioni dirette tra origine e destinazione.
  - **Strumento:** plotly.graph_objects.Scattergeo.[19]
  - **Metodologia:** Si tracciano linee o archi tra i centroidi (latitudine/longitudine) del Source_Country e del Target_Country. I dati aggregati (flow_counts) possono essere utilizzati per definire lo spessore o il colore della linea, rappresentando il volume di attacchi.[18]
- Visualizzazione 3: Diagrammi di Sankey
  Sebbene non sia una mappa geografica, un diagramma di Sankey è una delle visualizzazioni più efficaci per mostrare i flussi.20
  - **Strumento:** plotly.graph_objects.Sankey.[21]
  - **Metodologia:** Si definiscono i Source (es. Source_Country), Target (es. Target_Country) e Value (il volume di attacchi).[22] Il diagramma mostrerà chiaramente le proporzioni, ad esempio, come gli attacchi da un singolo paese si distribuiscono verso molteplici obiettivi.[20]

1.5 Tabella di Analisi dei Flussi: Top 15 Flussi di Attacco Nazione-Nazione Simulati

L'analisi geospaziale culmina in una sintesi che combina la frequenza (volume) con l'impatto (gravità) e l'efficacia (successo). La tabella seguente (ipotetica, basata sull'analisi descritta) rappresenta l'output finale desiderato da questa analisi.

Source_Country	Target_Country	Attack_Volume	Avg_Severity	Success_Rate

		(Conteggio)	(Media di Attack Severity)	(Percentuale di Outcome == 'Succeeded')
Nazione A	Nazione B	1,200	8.5	75%
Nazione C	Nazione B	950	4.2	30%
Nazione A	Nazione D	800	7.1	60%
Nazione E	Nazione F	750	9.0	85%
Nazione C	Nazione A	600	5.0	40%
...	...	...	...	...

Nota: La tabella sopra è un modello. I valori effettivi devono essere calcolati dal dataset.
Questa tabella fornisce un'intelligence molto più ricca del solo volume. Ad esempio, la rotta "Nazione C -> Nazione B" potrebbe rappresentare un alto volume di attacchi DDoS a bassa gravità e basso successo (media gravità 4.2, successo 30%). Al contrario, la rotta "Nazione E -> Nazione F" indica un problema più serio: un volume moderato di attacchi altamente gravi e molto efficaci (media gravità 9.0, successo 85%), suggerendo campagne di malware o infiltrazione mirate e di successo.[1]

Sezione 2: Valutazione Critica - L'Impatto Ineludibile dei Dati Sintetici (Generati da ChatGPT)

Una componente fondamentale dell'analisi di livello esperto è la valutazione critica della fonte dei dati. Questa sezione fornisce un contesto indispensabile per interpretare correttamente i risultati della Sezione 1.

2.1 L'Origine del Dataset: "A Synthetic Creation"

La descrizione del dataset su Kaggle è esplicita e deve essere il punto di partenza di ogni analisi: "This dataset is a synthetic creation, generated using ChatGPT to simulate realistic cybersecurity incidents".[1]

- **Implicazione di Primo Ordine:** I dati non sono reali. Non provengono da *honeypots*, *log di firewall*, *sistemi IDS/IPS* [23] o *report di incidenti* (come quelli visti in altri dataset, es. [24]). Sono interamente fittizi.
- **Implicazione di Secondo Ordine (Per l'Analisi Geospaziale):** L'analisi nazione-nazione condotta nella Sezione 1 *non* misura e *non può* misurare l'attività di attori statali (APT), gruppi di *hacktivisti* o reti criminali del mondo reale.
- **Implicazione di Terzo Ordine (La Vera Natura dell'Analisi):** Ciò che l'analisi geospaziale misura sono i **pattern statistici e i bias appresi da ChatGPT** durante il

suo addestramento.[25] Se l'analisi mostra che la "Nazione A" attacca frequentemente la "Nazione B", questo risultato non è un'indicazione di tensione geopolitica reale. È, invece, un'indicazione che i dati di addestramento di ChatGPT (composti da miliardi di pagine web, articoli di notizie, report sulla sicurezza, blog, ecc.) contengono una **forte associazione testuale** tra "Nazione A", "Nazione B" e "cyber attacco". L'analisi, quindi, si trasforma da un'analisi di *cybersecurity* a una *meta-analisi dell'IA*, che studia gli artefatti, i bias [28] e le potenziali "allucinazioni" [27] del modello generativo.

2.2 Limiti dell'Attribuzione IP nel Mondo Reale vs. Sintetico

L'analisi geospaziale basata su IP è intrinsecamente problematica, sia nel mondo reale che in questo contesto sintetico.

- **Problema del Mondo Reale:** Nel *threat intelligence* reale, l'attribuzione di un attacco a una nazione basandosi *esclusivamente* sull'indirizzo IP di origine è una pratica analitica debole, spesso errata, e considerata un errore da principiante.[29] Gli attori sofisticati (sia statali che criminali) mascherano sistematicamente la loro origine.[30] Le tecniche includono:
    - **VPN e Reti di Anonimizzazione:** Utilizzo di Virtual Private Network (VPN) [29] e nodi di uscita della rete Tor.[29]
    - **Proxy e Infrastrutture Compromesse:** Utilizzo di *residential proxy* (per apparire come un utente domestico legittimo) [31], server cloud compromessi [29] o altri dispositivi (IoT, router) parte di una botnet.[29]
    - IP Spoofing: Falsificazione dell'indirizzo IP di origine, comune negli attacchi DDoS.29
      Un analista professionista non affermerebbe mai "La Nazione A ha attaccato", ma piuttosto: "Il traffico di attacco è stato osservato originare da un indirizzo IP geolocalizzato nella Nazione A, che è un noto nodo di uscita Tor".31
- **Problema del Mondo Sintetico:** L'opacità del processo di generazione di ChatGPT [32] aggiunge un ulteriore livello di incertezza. Non abbiamo garanzie su come siano stati generati gli IP. Il modello ha generato indirizzi IP casuali dall'intero spazio IPv4 e poi ha assegnato loro un paese in base ai suoi bias? O ha prima deciso una coppia di nazioni (es. "Attacco da Nazione A a Nazione B") e poi ha generato indirizzi IP fittizi che *sembrano* appartenere a quei paesi? Questa mancanza di trasparenza rende ogni interpretazione geospaziale priva di un fondamento reale.

2.3 Valore Residuo: Come Utilizzare Correttamente Questo Dataset

Nonostante queste limitazioni critiche, il dataset ha un valore immenso se utilizzato per il suo scopo dichiarato: l'apprendimento.[1]

1. **Sandbox Metodologica:** Il dataset è uno strumento *eccellente* per l'apprendimento.[1] Permette di costruire e *testare* pipeline di analisi complesse (come l'arricchimento geospaziale della Sezione 1 o i modelli ML della Sezione 3) in un ambiente controllato.[33] Si evitano le sfide tipiche dei dati reali: problemi di privacy [33], formati di log disordinati e non strutturati, dati mancanti, e la necessità di gestire PII (Personally Identifiable Information).[35]
2. **Analisi dei Bias del Modello:** I risultati dell'analisi geospaziale sono un'opportunità per un progetto di ricerca diverso: studiare i bias intrinseci nei modelli LLM.[25] Quali nazioni ChatGPT etichetta come principali "attaccanti"? I risultati riflettono stereotipi geopolitici comuni diffusi dai media?

Conclusione Strategica: Si raccomanda di procedere con l'analisi della Sezione 1, ma di presentare i risultati non come "Un report sulla cyberwarfare globale", ma come "Un'analisi dei pattern geospaziali simulati nel Cyber Crimes Dataset generato da ChatGPT".

Sezione 3: Raccomandazioni Analitiche Aggiuntive (Analisi Supervisionata e Predittiva)

Sebbene l'analisi geospaziale sia la prospettiva richiesta, il dataset si presta a diverse altre analisi di *machine learning* supervisionato che hanno un valore pratico e formativo forse anche superiore.

3.1 Previsione 1 (Priorità Alta): Classificazione della Gravità dell'Attacco (Attack Severity)

- **Valore Operativo:** In un Security Operations Center (SOC) reale, gli analisti sono costantemente sommersi da un volume enorme di *alert*.[23] Molti di questi sono falsi positivi. Un modello di *machine learning* in grado di predire automaticamente la colonna Attack Severity [1] è uno strumento fondamentale. Permette al SOC di implementare un *triage* automatico, dando priorità agli alert "ad alta gravità" e ottimizzando i tempi di risposta (come Response Time nel dataset).[1]
- **Variabile Target:** Attack Severity.[1] Questa è una variabile categorica (probabilmente ordinale, es. Low, Medium, High) [41], rendendolo un problema di classificazione multi-classe.
- Feature Engineering (L'Approccio Esperto):
  Il successo di questo modello dipende da un'efficace feature engineering.42 Invece di inserire ciecamente tutte le colonne nel modello, un approccio esperto mappa le feature del dataset ai framework di valutazione del rischio del mondo reale, come quelli utilizzati da CISA o descritti in 43 e.43

    - Functional Impact (Impatto Funzionale) [43]: Caratterizzato da Industry e Target System.[1] Un attacco al settore "Finance" o a un "database" ha un impatto funzionale intrinsecamente più alto di un attacco a un "sito web" o a un "external user".[1]
    - Information Impact (Impatto Informativo) [43]: Caratterizzato direttamente dalla feature Data Compromised.[1] Questa è probabilmente la feature più predittiva dell'impatto.
    - Recoverability (Recuperabilità) [43]: Caratterizzato da Response Time e Attack Duration.[1]
    - Attack Characteristics (Caratteristiche dell'Attacco) [38]: Caratterizzato da Attack Type (es. Malware, DDoS), Security Tools Used (la presenza di difese) e User Role (un attacco a un 'admin' è più grave).[1]
- Modellazione e Interpretazione:
  Classificatori standard come Random Forest, XGBoost o Support Vector Machines (SVM) sono noti per le loro elevate prestazioni su dati tabellari di cybersecurity.44 Dopo l'addestramento, l'analisi della feature importance (es. SHAP values) 40 rivelerebbe cosa la simulazione di ChatGPT considera un fattore determinante per la gravità. È la quantità di dati (Data Compromised), il tipo di attacco (Attack Type) o il settore target (Industry)?

3.2 Tabella: Mappatura delle Feature per la Previsione della Gravità

Questa tabella illustra come un data scientist esperto di dominio tradurrebbe la conoscenza del dominio in feature ingegnerizzate per il modello.[42]

Dataset_Feature	**Conceptual_Framework_Feature**	**Engineering_Strategy (Esempio)**
Attack Type	Actor Characteristics / Attack Characteristics	One-Hot Encoding
Target System	Functional Impact / Importance of Target	One-Hot Encoding o Target Encoding
Industry	Functional Impact / Importance of Target	One-Hot Encoding o Target Encoding
Data Compromised	Information Impact / Economic Impact	Normalizzazione (es. Log-Transform, dato che è GB)
Attack Duration	Functional Impact / Duration	Normalizzazione (es. Min-Max Scaler)
Security Tools Used	(Fattore di mitigazione)	Binarizzazione o Conteggio (numero di strumenti)

User Role	Functional Impact	Encoding Ordinale (es. External=1, Employee=2, Admin=3)

Nota sulla Causalità: È fondamentale distinguere tra *previsione* e *valutazione*. Feature come Data Compromised e Response Time sono *risultati* di un attacco, non predittori disponibili *prima* che l'attacco abbia successo.

1. **Modello di Valutazione (Post-Incidente):** Utilizza tutte le feature per determinare la gravità di un incidente *concluso*.
2. **Modello Predittivo (Allerta in Tempo Reale):** Utilizza *solo* le feature disponibili al momento dell'allerta (es. Attack Type, Target System, Industry, Attacker IP) per predire la *potenziale* gravità, permettendo un triage proattivo.[23]

3.3 Previsione 2 (Priorità Media): Previsione dell'Esito (Outcome)

- **Valore Operativo:** Risponde a una domanda chiave per la gestione della sicurezza: "Dato un attacco con queste caratteristiche e le nostre difese attuali, avrà successo?"
- **Variabile Target:** Outcome (Succeeded/Failed).[1] Si tratta di un problema di classificazione binaria.
- **Test di Coerenza della Simulazione:** Questo modello serve anche come un potente *test di coerenza* per il dataset sintetico. Logicamente, le feature più importanti per predire il successo o il fallimento dovrebbero essere Security Tools Used, Response Time e Mitigation Method.[1] Ci si aspetta una forte correlazione negativa tra la presenza di difese (es. "Firewall", "IDS") e il successo dell'attacco. Se un modello di machine learning non riesce a trovare questa relazione, significa che i dati sintetici sono incoerenti o rumorosi, e la logica interna della simulazione è difettosa.[28]

3.4 Tabella: Confronto Modelli per Classificazione Attack Severity

Per un progetto completo, si raccomanda un benchmark di diversi modelli per identificare il più performante.

Modello	**Accuracy (Media)**	**Precision (Macro)**	**Recall (Macro)**	**F1-Score (Macro)**	**Note**
Logistic Regression [45]	0.75	0.72	0.71	0.71	Veloce, interpretabile, buon baseline.
Random Forest [44]	0.88	0.86	0.85	0.85	Ottimo per dati tabellari, robusto, interpretabile

					(feature importance).
XGBoost [45]	0.90	0.89	0.88	0.88	Spesso il migliore in performance, richiede tuning degli iperparametri.
MLP (Neural Network) [46]	0.89	0.87	0.87	0.87	Potenziale elevato, ma rischio di overfitting su dati non complessi.

Nota: I valori sono ipotetici. L'uso di metriche "Macro" (es. F1-Score Macro) è fondamentale, poiché le classi di gravità (es. 'Critica') sono probabilmente molto meno numerose delle classi 'Bassa', un problema comune di *class imbalance* nei dati di sicurezza.[32]

Sezione 4: Raccomandazioni Analitiche Aggiuntive (Analisi delle Serie Temporali)

La colonna Timestamp [1] è una delle risorse più preziose del dataset, ma spesso trascurata. Permette di spostare l'analisi da una visione statica ("quanti attacchi") a una dinamica ("quando e con quale frequenza avvengono gli attacchi").

4.1 Analisi Esplorativa (EDA) Temporale

- **Pre-processing:** Il primo passo è la conversione della colonna Timestamp da stringa a un oggetto datetime di pandas [47], impostandolo preferibilmente come indice del DataFrame (DatetimeIndex).[49]
- **Analisi dei Pattern:** Aggregare i conteggi degli attacchi nel tempo.
    - **Resampling:** Usare pandas.resample() per aggregare il numero di attacchi per ora (.resample('H').size()), giorno ('D') [50] o settimana ('W').
    - **Pattern Ciclici:** Estrarre componenti temporali come l'ora del giorno (.dt.hour) o il giorno della settimana (.dt.weekday).[49]
- **Domande da porsi:** La simulazione di ChatGPT mostra bias?
    - Gli attacchi si concentrano nelle "ore lavorative" (es. 9:00-17:00)?
    - C'è un calo durante i fine settimana?[51]
    - Nel mondo reale, l'attività delle botnet è 24/7, mentre gli attacchi mirati (APT)

possono essere programmati per avvenire fuori orario lavorativo per evitare il rilevamento. I pattern trovati in questo dataset riveleranno, ancora una volta, i presupposti del modello generativo.

- **Visualizzazione:** Utilizzare grafici a linee (matplotlib.pyplot [52]) per mostrare i trend e heatmap (seaborn.heatmap) per visualizzare l'intensità degli attacchi per x [Ora del Giorno].

4.2 Forecasting dei Volumi di Incidente (Previsione)

- **Valore Operativo:** Prevedere il volume di attacchi futuri (es. "quanti attacchi ci aspettiamo domani?") è fondamentale per la gestione di un SOC. Permette di allocare le risorse, pianificare i turni del personale e scalare l'infrastruttura di sicurezza in anticipo rispetto ai picchi previsti.
- **Metodologia 1: Modelli Statistici (ARIMA/SARIMA)**
  - **Uso:** I modelli ARIMA (Autoregressive Integrated Moving Average) sono un approccio statistico classico per il forecasting di serie temporali.[53]
  - **Passaggi:** (1) Creare la serie temporale (es. conteggio attacchi al giorno). (2) Controllare la stazionarietà (es. ADF test) ed eseguire la differenziazione (la 'I' in ARIMA) se necessario.[53] (3) Usare i grafici ACF/PACF per determinare i parametri (p, d, q).
  - Se l'analisi EDA rivela una forte stagionalità settimanale (un pattern che si ripete ogni 7 giorni), un modello SARIMA (Seasonal ARIMA) è più appropriato e probabilmente più performante.[54]
- **Metodologia 2: Modelli di Deep Learning (LSTM)**
  - **Uso:** Le reti neurali Long Short-Term Memory (LSTM) sono una forma di RNN (Recurrent Neural Network) [56] particolarmente adatte a catturare pattern temporali complessi e non lineari, superando i problemi di memoria a breve termine (vanishing gradient) degli RNN tradizionali.[57]
  - **Preparazione Dati:** Le LSTM richiedono una trasformazione della serie temporale in un problema supervisionato. Questo viene fatto creando "finestre" (o *sequences*).[58] Ad esempio, si usano n_past=10 time step (es. 10 giorni) come feature (X) per predire n_future=1 time step (es. il giorno successivo) come target (y).[50]
  - Questo dataset offre un'eccellente opportunità per un progetto di confronto: implementare sia un modello ARIMA/SARIMA [53] che un modello LSTM [58] e valutare quale dei due produce previsioni più accurate per questo specifico flusso di dati sintetici.

Sezione 5: Raccomandazioni Analitiche Aggiuntive

(Analisi Non Supervisionata)

Questa classe di analisi è progettata per scoprire pattern nascosti senza fare affidamento su etichette predefinite come Attack Type o Attack Severity. Si tratta di *threat hunting* (caccia alle minacce) [59] e di scoperta di profili.

5.1 Rilevamento di Anomalie (Anomaly Detection)

- **Obiettivo:** Identificare gli incidenti *più strani*, *inaspettati* o *aberranti* nel dataset.[60] Questi "outlier" [61] non sono necessariamente gli attacchi più "gravi", ma quelli che deviano maggiormente dalla norma, e potrebbero rappresentare minacce nuove o emergenti (unknown unknowns).[59]
- **Distinzione Critica (Anomalia vs. Gravità):**
    - *Esempio 1 (Alta Gravità, Non Anomalo):* Un attacco "Ransomware" [1] che compromette 10GB di dati e riceve una Attack Severity di 9/10 potrebbe essere molto *grave*, ma se ci sono 500 attacchi simili, non è *anomalo*. È un pattern noto.
    - *Esempio 2 (Bassa Gravità, Molto Anomalo):* Un attacco "Phishing" [1] con un Attack Duration [1] di 5000 minuti e un Response Time [1] di 1 minuto è *estremamente strano*. È un'anomalia statistica. Potrebbe essere un errore nei dati sintetici o, nel mondo reale, un tipo di attacco persistente non etichettato correttamente che un analista dovrebbe indagare immediatamente.[60]
- **Modelli:**
    - **Isolation Forest:** Un modello basato su alberi, eccellente per dati tabellari ad alta dimensionalità. Isola le anomalie velocemente.[62]
    - **Autoencoder (Deep Learning):** Un modello di rete neurale addestrato a "comprimere" (encoder) e "ricostruire" (decoder) i dati.[63] Il modello impara bene a ricostruire gli attacchi *normali* e *comuni*. Quando viene presentato un attacco *anomalo*, l'autoencoder fallisce nel ricostruirlo accuratamente, risultando in un "errore di ricostruzione" elevato. Gli incidenti con l'errore più alto sono le anomalie più significative.

5.2 Clustering: Scoperta di Profilo di Attacco (Attack Profiling)

- **Obiettivo:** Invece di usare le etichette fornite (Attack Type), il clustering permette di scoprire se i dati stessi si raggruppano in "archetipi" di attacco basati sulle loro caratteristiche.[59]
- **Algoritmi:** K-Means (per cluster sferici e un numero k predefinito) [65] o DBSCAN (basato sulla densità, ottimo per trovare rumore/anomalie).[65]

- **Feature:** L'analisi dovrebbe essere eseguita su un set di feature numeriche e categoriche codificate (es. Attack Duration, Data Compromised, Target System (codificato), Industry (codificato)).[1]
- **Interpretazione dei Cluster:** Dopo aver eseguito l'algoritmo (es. K-Means con $k=5$), il passo successivo è analizzare il "centroide" o i membri medi di ciascun cluster. Si potrebbero scoprire profili che il dataset non etichetta esplicitamente [64]:
    - *Cluster 0:* "Attacchi veloci, a basso impatto, automatizzati" (es. alta frequenza, bassa durata, basso Data Compromised).
    - *Cluster 1:* "Attacchi lenti e persistenti, mirati al settore finanziario" (es. lunga durata, alto Data Compromised, Industry == 'Finance').
    - Cluster 2: "Attacchi DDoS su larga scala" (es. durata estremamente elevata, Target System == 'Server', basso Data Compromised).
      Questo approccio permette di identificare TTPs (Tattiche, Tecniche e Procedure) 59 emergenti dalla simulazione stessa.

Sezione 6: Conclusione e Raccomandazioni Strategiche

Il "Cyber Crimes Dataset" (shakirul09) [1] è uno strumento di *apprendimento* eccezionale. La sua natura sintetica, generata da ChatGPT [1], è contemporaneamente il suo più grande limite e la sua più grande forza.

1. **Azione 1: Eseguire l'Analisi Geospaziale (Priorità Utente).** Si raccomanda di procedere con l'analisi nazione-nazione (Sezione 1) come un esercizio metodologico robusto. Questo progetto è ideale per padroneggiare l'arricchimento IP di massa (usando geoip2 [2]) e le tecniche di visualizzazione avanzata dei flussi (usando plotly per *Choropleth* [16], *Scattergeo* [19] e *Sankey* [21]).
2. **Azione 2: Contestualizzare Criticamente i Risultati.** È *imperativo* che i risultati dell'analisi geospaziale siano inquadrati correttamente (Sezione 2). Questa non è *threat intelligence* del mondo reale. È un'*analisi del simulacro*. I risultati riflettono i bias statistici del modello LLM che ha generato i dati [25] e non tengono conto delle complessità del mondo reale nell'attribuzione IP (es. VPN, proxy).[29] Qualsiasi report su questa analisi deve includere questa *caveat* in modo prominente.
3. **Azione 3: Espandere l'Analisi (Valore Aggiunto).** L'analisi di maggior valore "pronta per l'industria" che questo dataset pulito e sintetico [33] consente, risiede nelle analisi supervisionate e non supervisionate. Si raccomanda di espandere il progetto per includere:
    - **Predizione della Gravità (Sezione 3):** Per simulare il *triage* e la *prioritizzazione degli alert* in un SOC, un'abilità fondamentale.[40]
    - **Forecasting dei Volumi (Sezione 4):** Per simulare la *gestione operativa* delle risorse di sicurezza, confrontando modelli statistici (ARIMA [53]) e di deep learning

(LSTM [57]).

- **Rilevamento delle Anomalie (Sezione 5):** Per simulare il *threat hunting* proattivo, andando oltre le etichette fornite per trovare gli "unknown unknowns".[60]

In sintesi, questo dataset è una "palestra" perfetta.[1] Permette di allenare i "muscoli" analitici (le metodologie di data science) su macchine pulite, sicure e funzionanti (i dati sintetici), preparando l'analista ad affrontare il caos, il rumore e la complessità dei "log del mondo reale". L'analisi geospaziale è un eccellente punto di partenza, ma è solo il primo passo di un percorso di data science molto più ricco che questo dataset può supportare pienamente.

Bibliografia

1. Cyber Crimes Dataset - Kaggle, accesso eseguito il giorno novembre 13, 2025, https://www.kaggle.com/datasets/shakirul09/cyber-crimes-dataset
2. Data Enrichment Series - Part 1 - Bulk IP Triage with Python and MaxMind GeoIP - KC7, accesso eseguito il giorno novembre 13, 2025, https://kc7cyber.com/blog/data-enrichment-series-part-1-bulk-ip-triage-with-python-and-maxmind-geoip
3. Visualizing Geo IP Information using Python - Recon InfoSec, accesso eseguito il giorno novembre 13, 2025, https://blog.reconinfosec.com/visualizing-geo-ip-information-in-python
4. IP2Location™ LITE IP-COUNTRY Database - Kaggle, accesso eseguito il giorno novembre 13, 2025, https://www.kaggle.com/datasets/ip2location/ip2location-lite-ip-country-database
5. Global IP Dataset by Location 2023 - Kaggle, accesso eseguito il giorno novembre 13, 2025, https://www.kaggle.com/datasets/joebeachcapital/global-ip-dataset-by-location-2023
6. How to Track the Location of an IP Address using Python - KDnuggets, accesso eseguito il giorno novembre 13, 2025, https://www.kdnuggets.com/2023/01/track-location-ip-address-python.html
7. tomas-net/ip2geotools: Simple tool for getting geolocation information on given IP address from various geolocation databases. - GitHub, accesso eseguito il giorno novembre 13, 2025, https://github.com/tomas-net/ip2geotools
8. How to Track IP Address Using Python? - Analytics Vidhya, accesso eseguito il giorno novembre 13, 2025, https://www.analyticsvidhya.com/blog/2024/06/track-ip-address-using-python/
9. Using Python to Convert IP Addresses Into Location Data - IPstack, accesso eseguito il giorno novembre 13, 2025, https://ipstack.com/blog/using-python-to-convert-ip-addresses-into-location-data
10. IP Geolocation analysis in Python made simple - WhoisXML API, accesso eseguito il giorno novembre 13, 2025, https://ip-geolocation.whoisxmlapi.com/blog/ip-geolocation-analysis-in-python-

made-simple
11. Pandas: fastest way to resolve IP to country - python - Stack Overflow, accesso eseguito il giorno novembre 13, 2025, https://stackoverflow.com/questions/40211314/pandas-fastest-way-to-resolve-ip-to-country
12. maxmind/GeoIP2-python: Python code for GeoIP2 webservice client and database reader - GitHub, accesso eseguito il giorno novembre 13, 2025, https://github.com/maxmind/GeoIP2-python
13. andrusha/pandas-maxminddb: Fast geolocation library for Pandas dataframes written in Rust - GitHub, accesso eseguito il giorno novembre 13, 2025, https://github.com/andrusha/pandas-maxminddb
14. Maps in Python - Plotly, accesso eseguito il giorno novembre 13, 2025, https://plotly.com/python/maps/
15. Plotting Location of Attacks - Kaggle, accesso eseguito il giorno novembre 13, 2025, https://www.kaggle.com/code/jscearce5/plotting-location-of-attacks
16. Choropleth maps in Python - Plotly, accesso eseguito il giorno novembre 13, 2025, https://plotly.com/python/choropleth-maps/
17. Choropleth Maps using Plotly in Python - GeeksforGeeks, accesso eseguito il giorno novembre 13, 2025, https://www.geeksforgeeks.org/python/choropleth-maps-using-plotly-in-python/
18. Visualizing Network Optimization Model Results Using Python - OR & Data Science Stories, accesso eseguito il giorno novembre 13, 2025, https://emrahcimren.github.io/visualization/Visualizing-Network-Optimization-Model-Results-using-Python/
19. Lines on maps in Python - Plotly, accesso eseguito il giorno novembre 13, 2025, https://plotly.com/python/lines-on-maps/
20. Here's How to use Sankey Diagrams for Data Visualization - Analytics Vidhya, accesso eseguito il giorno novembre 13, 2025, https://www.analyticsvidhya.com/blog/2021/11/visualize-data-using-sankey-diagram/
21. Sankey diagram in Python - Plotly, accesso eseguito il giorno novembre 13, 2025, https://plotly.com/python/sankey-diagram/
22. Understanding Plotly Sankey Diagrams | by Tom Welsh - Medium, accesso eseguito il giorno novembre 13, 2025, https://medium.com/@twelsh37/understanding-plotly-sankey-charting-3ee263a81549
23. Cyber Attack EDA - Kaggle, accesso eseguito il giorno novembre 13, 2025, https://www.kaggle.com/code/saadatkhalid/cyber-attack-eda
24. Cyber Crime dataset - Kaggle, accesso eseguito il giorno novembre 13, 2025, https://www.kaggle.com/datasets/mdriajuliislam/cybercrime
25. ChatGPT Security Risks: All You Need to Know - SentinelOne, accesso eseguito il giorno novembre 13, 2025, https://www.sentinelone.com/cybersecurity-101/data-and-ai/chatgpt-security-risks/
26. Security Analysis of ChatGPT: Threats and Privacy Risks - arXiv, accesso eseguito

il giorno novembre 13, 2025, https://arxiv.org/html/2508.09426v1
27. The Limitations and Ethical Considerations of ChatGPT | Data Intelligence - MIT Press Direct, accesso eseguito il giorno novembre 13, 2025, https://direct.mit.edu/dint/article/6/1/201/118839/The-Limitations-and-Ethical-Considerations-of
28. accesso eseguito il giorno novembre 13, 2025, https://www.forbes.com/councils/forbestechcouncil/2025/08/21/using-synthetic-data-consider-19-pros-and-cons-from-tech-leaders/#:~:text=By%20simulating%20realistic%20scenarios%2C%20it,reality%20and%20undermine%20model%20performance.
29. Geo-blocking in context: Realities, risks and recommendations | Cyber.gov.au, accesso eseguito il giorno novembre 13, 2025, https://www.cyber.gov.au/business-government/protecting-devices-systems/hardening-systems-applications/network-hardening/geo-blocking-in-context-realities-risks-recommendations
30. accesso eseguito il giorno novembre 13, 2025, https://www.cyber.gov.au/business-government/protecting-devices-systems/hardening-systems-applications/network-hardening/geo-blocking-in-context-realities-risks-recommendations#:~:text=While%20IP%20addresses%2C%20domain%20names,conceal%20their%20identities%20and%20location.
31. Rethinking Identity Threat Detection: Don't Rely on IP Geolocation - Obsidian Security, accesso eseguito il giorno novembre 13, 2025, https://www.obsidiansecurity.com/blog/rethinking-identity-threat-detection-the-ip-geolocation
32. Leveraging data analytics to revolutionize cybersecurity with machine learning and deep learning - PMC - PubMed Central, accesso eseguito il giorno novembre 13, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC12397323/
33. Synthetic Data Generation in Cybersecurity: A Comparative Analysis - arXiv, accesso eseguito il giorno novembre 13, 2025, https://arxiv.org/html/2410.16326v1
34. Using Privacy Preserving Synthetic Data to Enhance Cyber Security Ops for DHS, accesso eseguito il giorno novembre 13, 2025, https://www.betterdata.ai/blogs/using-privacy-preserving-synthetic-data-to-enhance-cyber-security-ops-for-dhs
35. The use of Synthetic Data to protect from Cybersecurity threats caused by Unstructured Content in the Generative AI Era | by Dr. Shweta Shah | Medium, accesso eseguito il giorno novembre 13, 2025, https://medium.com/@drshwetashah/the-use-of-synthetic-data-to-protect-from-cybersecurity-threats-caused-by-unstructured-content-in-e0fc3243b3ab
36. What is synthetic data — and how can it help you competitively? - MIT Sloan, accesso eseguito il giorno novembre 13, 2025, https://mitsloan.mit.edu/ideas-made-to-matter/what-synthetic-data-and-how-can-it-help-you-competitively
37. The benefits of using synthetic data in cybersecurity - Syntheticus, accesso eseguito il giorno novembre 13, 2025, https://syntheticus.ai/blog/the-benefits-of-using-synthetic-data-in-cybersecurity

38. Cyber Attacker Profiling for Risk Analysis Based on Machine Learning - MDPI, accesso eseguito il giorno novembre 13, 2025, https://www.mdpi.com/1424-8220/23/4/2028
39. Cyber Security Attacks - Kaggle, accesso eseguito il giorno novembre 13, 2025, https://www.kaggle.com/datasets/teamincribo/cyber-security-attacks
40. AI-Driven Real-Time Severity Prediction for Cyber Attacks using Machine Learning, accesso eseguito il giorno novembre 13, 2025, https://ieeexplore.ieee.org/document/11081230/
41. Cybersecurity Threat and Awareness Program Dataset - Kaggle, accesso eseguito il giorno novembre 13, 2025, https://www.kaggle.com/datasets/datasetengineer/cybersecurity-threat-and-awareness-program-dataset
42. Data Preprocessing and Feature Engineering for Cyber Threat Detection - ResearchGate, accesso eseguito il giorno novembre 13, 2025, https://www.researchgate.net/publication/379078896_Data_Preprocessing_and_Feature_Engineering_for_Cyber_Threat_Detection
43. A Cyberattack Severity Classification Framework for the Republic of ..., accesso eseguito il giorno novembre 13, 2025, https://www.csis.org/analysis/cyberattack-severity-classification-framework-republic-korea
44. Machine learning classifier-based detection of cyber-attack on power system: Comparative analysis - IEEE Xplore, accesso eseguito il giorno novembre 13, 2025, https://ieeexplore.ieee.org/document/10069603/
45. Machine Learning-Based Methodologies for Cyber-Attacks and Network Traffic Monitoring: A Review and Insights - MDPI, accesso eseguito il giorno novembre 13, 2025, https://www.mdpi.com/2078-2489/15/11/741
46. An efficient cyber-attack detection and classification in IoT networks with high-dimensional feature set using Levenberg-Marquardt optimized feedforward neural network | PLOS One, accesso eseguito il giorno novembre 13, 2025, https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0333899
47. CyberSecurityDataset_EDA_Dat, accesso eseguito il giorno novembre 13, 2025, https://www.kaggle.com/code/csdataset/cybersecuritydataset-eda-datapreparation
48. Tutorial: Time Series Analysis with Pandas - Dataquest, accesso eseguito il giorno novembre 13, 2025, https://www.dataquest.io/blog/tutorial-time-series-analysis-with-pandas/
49. How to handle time series data with ease — pandas 2.3.3 documentation - PyData |, accesso eseguito il giorno novembre 13, 2025, https://pandas.pydata.org/docs/getting_started/intro_tutorials/09_timeseries.html
50. Time series forecasting using LSTM - Kaggle, accesso eseguito il giorno novembre 13, 2025, https://www.kaggle.com/code/gurpreetmohaar/time-series-forecasting-using-lstm
51. Time Series Visualization of Network Activity - Kaggle, accesso eseguito il giorno novembre 13, 2025,

https://www.kaggle.com/code/ernie55ernie/time-series-visualization-of-network-activity
52. Guide to Time-Series Analysis in Python | Tiger Data, accesso eseguito il giorno novembre 13, 2025, https://www.tigerdata.com/blog/how-to-work-with-time-series-in-python
53. Time Series Forecasting - ARIMA, LSTM, Prophet - Kaggle, accesso eseguito il giorno novembre 13, 2025, https://www.kaggle.com/code/cdabakoglu/time-series-forecasting-arima-lstm-prophet
54. Time Series Forecasting - ARIMA - Kaggle, accesso eseguito il giorno novembre 13, 2025, https://www.kaggle.com/code/mechatronixs/time-series-forecasting-arima
55. Time Series Forecast: A comprehensive Guide - Kaggle, accesso eseguito il giorno novembre 13, 2025, https://www.kaggle.com/code/ankumagawa/time-series-forecast-a-comprehensive-guide
56. Analyzing and Predicting Cyber Hacking with Time Series Models - International Journal of Research in Engineering, Science and Management - IJRESM, accesso eseguito il giorno novembre 13, 2025, https://journal.ijresm.com/index.php/ijresm/article/download/5/2/4
57. Intro to LSTM 🗓️ Time Series Forecasting - Kaggle, accesso eseguito il giorno novembre 13, 2025, https://www.kaggle.com/code/azminetoushikwasi/intro-to-lstm-time-series-forecasting
58. Predicting DDoSAttacks with LSTM Time-Series model - Kaggle, accesso eseguito il giorno novembre 13, 2025, https://www.kaggle.com/code/asmahwimli/predicting-ddosattacks-with-lstm-time-series-model
59. Clustering for Threat Detection: Machine Learning in Cybersecurity - Uptycs, accesso eseguito il giorno novembre 13, 2025, https://www.uptycs.com/blog/threat-research-report-team/machine-learning-in-cybersecurity
60. What is Anomaly Detection? - Wiz, accesso eseguito il giorno novembre 13, 2025, https://www.wiz.io/academy/anomaly-detection
61. What Is Anomaly Detection? - CrowdStrike, accesso eseguito il giorno novembre 13, 2025, https://www.crowdstrike.com/en-us/cybersecurity-101/next-gen-siem/anomaly-detection/
62. Deep Learning-based Anomaly Detection and Log Analysis for Computer Networks - arXiv, accesso eseguito il giorno novembre 13, 2025, https://arxiv.org/abs/2407.05639
63. Anomaly Detection within Machine Learning on Logs - SANS Institute, accesso eseguito il giorno novembre 13, 2025, https://www.sans.org/webcasts/anomaly-detection-within-machine-learning-logs

64. Karan-D-Software/Machine-Learning-Network-Security: We perform clustering analysis on the CICIDS2017 dataset to identify patterns and group similar network traffic for cybersecurity insights. - GitHub, accesso eseguito il giorno novembre 13, 2025, https://github.com/Karan-D-Software/Machine-Learning-Network-Security
65. A SURVEY ON THE USE OF DATA CLUSTERING FOR INTRUSION DETECTION SYSTEM IN CYBERSECURITY - PMC - NIH, accesso eseguito il giorno novembre 13, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC8289996/
66. Detecting Cyber Threats in UWF-ZeekDataFall22 Using K-Means Clustering in the Big Data Environment - MDPI, accesso eseguito il giorno novembre 13, 2025, https://www.mdpi.com/1999-5903/17/6/267
67. Threat hunting in large datasets by clustering security events - Cisco Talos Blog, accesso eseguito il giorno novembre 13, 2025, https://blog.talosintelligence.com/threat-hunting-in-large-datasets-by/